< preprint >

The economics of English as a global language: Evidence from 2.0 mln subjects suggests that an economic model substantially accounts for country differences in English proficiency

Roeland van Hout[1] and Frans van der Slik[1, 2]

Abstract

We have analyzed Education First's Standard English Test (SET) scores coming from 2 mln learners of English in 110 countries and regions world-wide. There is a significant high correlation between the country SET scores and their economic Human Capital Index. We introduce a quantitative linguistic distance measure based on comparing words (lexicon), constructions (morphology), and sounds (phonology). Linguistic distances were computed between English, the second language, and the main mother tongues of the counties involved. We distinguished 61 mother tongues. The linguistic distance measure has a significant high correlation with the country SET scores as well. The larger the distance the lower the country SET score. The third explanatory variable we apply is the official status of English in the 110 countries. We relate our variables to the economics of language model from Chiswick and Miller (1995) that they developed for a different group of English language learners, immersion learners, i.e. immigrants in English speaking countries.

We applied mixed regression analyses including the 61 country language as a random factor with the Human Capital Index, the status of English and linguistic distance as fixed factors. The country scores are largely based on non-immersion learners of English. All three explanatory factors play an essential role in the final regression model that accounts for 51% of the total variance between countries. These three variables fit the economics of language model as they obviously relate to its three pivotal concepts: economic incentives, exposure to English, and efficiency.

[1] Centre for Language Studies, Radboud University, Nijmegen, the Netherlands

[2] Research Focus Area: Understanding and Processing Language in Complex Settings, North-West University, South Africa

The economics of English as a global language: Evidence from 2.0 mln subjects suggests that an economic model substantially accounts for country differences in English proficiency

Introduction

English is used as a lingua franca in international business communication, tourism, and in international scientific research all over the world. In many countries, English is included in the curriculum of secondary or even primary education. Around one billion people have learned English as a second language while more than 370 million people are using English as their first language (Ethnologue, 24th edition). English as a global language has a profound effect on many people's lives. Learning English in a non-English spoken environment augments career and income opportunities (see Tibus & Bendulo, 2018). Consequently, the English language testing industry has become a billion dollar enterprise (Azoth Analytics, 2019). Testing companies such as ETS, Cambridge, Education First, and countless others offer English courses and certificates with substantial civic prospects.

Not all learners of English are equally successful. Education First (2021), henceforth EF, has reported systematic differences between a large number of countries and regions. Dutch learners, for example, were found to score highest on EF's Standard English Test (SET) for some consecutive years now, while learners from other Germanic speaking countries such as Austria, Denmark and Norway turn up in the top level too. The only non-European country in EF's current top-5 is Singapore, a country with strong former colonial bounds with the UK (Education First, 2021), where English is one of the official languages and the primary medium of communication in business and trade (World Atlas, 2022). Despite some notable exceptions, learners from African and other Asian countries fare less well in English, however.

Various explanations can be put forward for the observed differences in average English proficiency scores between countries. According to the World Economic Forum (2022), the historical trade links between European countries and the UK and the fact that English along German and French is one of the so-called international 'working languages' can be hold responsible for the high rankings of speakers of European countries. A more encompassing explanation links the country differences to the level of economic development. Education First (2012, p. 12) reports a high correlation between countries' SET scores and the Human Capital Index (HCI; see Kraay, 2018). This index measures countries' success in mobilizing the economic and professional potential of their citizens. Chiswick and Miller (2014) outline the economics of language theory in which language skills among immigrants and native-born linguistic minorities are defined as a form of human capital satisfying the requirements of being productive, costly to produce, and embodied in the person. Chiswick and Miller (2014: 86) draw a unambiguous, economic conclusion: "Among immigrants, other variables being the same, earnings are greater for those more proficient in English." The foundations of the economics of language approach can be found in Chiswick and Miller (1995) where they investigate the determinants of immigrants' fluency in the dominant destination language, English. These determinants of proficiency relate to general concepts, indicated as the three E's: Economic incentives, Exposure, and Efficiency. Chiswick and Miller (2005) observed a substantial impact of linguistic distance on proficiency in the dominant language. The larger

the distance, the lower the proficiency. This determinant belongs to the concept of efficiency as it facilitates or hinders achieving target-language proficiency.

Interestingly, Takeno and Moritoshi (2018) suggest that linguistic dissimilarity or distance as operationalized by Chiswick and Miller (2005) might explain why Japanese learners score low on the SET. These Japanese learners are not immigrants, but non-immersion learners from Japan. We want to take up their suggestion, but there are some rather fundamental problems with the linguistic distance measure of Chiswick and Miller (1995). Van der Slik (2010) argued that the Chiswick and Miller (2005) linguistic distance measure suffers from severe validity problems. Basically, this measure takes into account the difficulty American learners of a foreign language subjectively report in learning that foreign language. A critical assumption underlying the validity of this particular measure of linguistic distance is that the distance between, for example, English and Japanese is the same as the distance between Japanese and English; that is, it is equally difficult for Americans to learn Japanese as it is for Japanese speakers to learn English. This assumption appears to be untenable due to asymmetric intelligibilities between languages (see van der Slik, 2010).

In earlier research, we (Schepens, 2015; Schepens, Van der Slik, & Van Hout, 2013, forthcoming; Schepens, Van Hout, & Jaeger, 2020; Van der Slik, 2010; Van der Slik, Van Hout, & Schepens, 2015, 2019) calculated dissimilarity or distance measures between Dutch and other languages following an algorithmic linguistic approach. First, we used a lexical distance measure that is based on the proportion of cognates two Indo-European languages share. The less overlap, the more distant the two Indo-European languages are. An obvious limitation of the lexical distance measure is that it is only applicable to languages of the Indo-European language family. Languages of other families do not have cognates with English. We additionally developed two more versatile, asymmetric measures, i.e., a morphological and a phonological distance measure that can be deployed to all language pairs, from whatever language family. The morphological measure scales the degree of morphological complexity of a specific language in relation to a target language, English for instance. The phonological measure takes the number of phonological features not present in a specific language compared to a target language. The operationalization of the three linguistic distance measures is explained in more detail in the method section (see also Schepens et al. forthcoming). We successfully investigated language proficiency levels in learning Dutch as a second language in relation to the linguistic distances of the mother tongue (L1) of adult language learners, i.e. an immersion context.

Our data is on English proficiency in, basically, a non-immigration context, meaning that we are dealing with the economics of English as a global language. Consequently, we need to link the three E's as theoretical concepts to concrete determinants that can be investigated and tested. The concept of economic incentives might include the increment in annual earnings because of a level of language proficiency or fluency for the individual learner, but in our case we prefer to interpret this concept as the educational context in which learners are being offered facilities and opportunities to learn, including learning English in their country of origin. The more a country is involved in global economic developments, the more it will invest in education and in teaching English, the language of international trade and communication. EF (2021) reports high correlations (.65 – .70) over the years between their index and economy-based country indices like human capital (World Bank, 2020), productivity capacity (UN Conference of Trade and Development, 2020), global talent

competitiveness (Lanvin & Monteiro, 2020), and global innovation indices (Dutta, Lanvin, & Wunsch-Vincent, 2020). All these measures have to do with economic potential. In the present study we selected the Human Capital Index (HCI). Other predictors might have been chosen but in our view the HCI offers the most comprehensive theoretical underpinning of a causal relationship between a country's investment in its citizens' resources and their educational outcomes, including learning English. Moreover, these economy-based measures are highly overlapping. We calculated their correlations for our SET data and they vary between .76 and .84.

The second factor of Chiswick and Miller (1995) is the amount of exposure and input. In a non-immigrant context, this factor defines the presence of English in the countries involved. In many non-European countries, English obtained a strong position because of their colonial past or because of intensive trading with Great Britain. These historical links often have led to the establishment of English as one of the official languages, next to one or more indigenous languages, or to English as a medium of instruction in schooling. In this way, English has become one of the official languages in many countries, We defined the status of English (SE) a potential determinant, being positive (an official status) or not (not an official status). This official presence may have multiple causes and various contexts, but its consequence is the exposure to and input of English in daily affairs and in education.

Their third concept, efficiency, refers to the extent to which exposure and learning in fact produce language proficiency. Degree of schooling is covered by our first two factors, having the consequence that efficiency primarily relates to our linguistic distance determinant. We have shown that earlier acquired knowledge (L1) constraints the acquisition of new knowledge (an additional language) (Schepens et al. forthcoming). Distance or dissimilarity is a restricting factor, because of biological constraints in cognitive resources, when language learners start later at older ages, but, in particular, because of the gap between existing and new language knowledge. In the present study, we want to use the approach we developed by testing if linguistic distance adds to explaining the observed country differences in the SET English proficiency scores.

Given our aim to investigate the contributions of our three determinants, human capital, status of English, and linguistic distance, we need to address the following questions:

1) *How can we operationalize linguistic distance as a determinant?*
   We propose to use three linguistic dissimilarity measures related to words, constructions and sounds and to integrate them in one overall quantitative index.
2) *What is the correlation of our three core determinants with the SET country scores?*
   Next to linguistic distance (LD), we distinguish status of English (SE) and the Human Capital Index (HCI).
3) *How do these determinants emerge when they are part of an overall analysis?*
   We will carry out linear mixed regressions to investigate the contributions of the three determinants to English proficiency, with language as a random effect. The criterion variable is the SET country score. We will also investigate relevant interactions in our data, to be sure what the precise contribution of each factor is.

Method

In the present study, we make use of the data published by Education in 2021. It comprises of aggregated EF SET scores of 2.0 million language test takers in 112 countries or regions. The SET can be taken at: https://www.efset.org/ef-set-50/. The description of the SET and its reliability can be found on https://www.efset.org/about. The SET consists of a reading and a listening test section. Their reliabilities are very high (.95 and .94, respectively; Education First, 2015b). In addition, external validity turned out to be satisfactory, as the correlations with IELTS and TOEFL iBT individual test scores were well above .50, as reported in Education First (2015a, 2015b).

The R codes and data used in the present study (Van Hout & Van der Slik, 20xx) are freely available at the Open Science Framework (https://osf.io/info).

Completion time of each part is limited to 25 minutes. The test has a computer-guided adaptive Multi-Stage Testing approach (EF, 2014). Although the test administration is too complex to be described here in a few lines, the essence of the SET test can be resumed as follows. A test taker's initial response to test items of medium difficulty will set the automated path to subsequent more difficult or less difficult test items/modules that are supposed to reflect a test taker's ability with increasing accuracy. The difficulty of test items was predetermined by means of Rasch analyses of test results of approximately 37,000 examinees. For a comprehensive outline of the test specifications we refer to EF (2014).

The SET is aimed at test takers that have at least a high school degree. After completing the test, test takers are invited to provide minimal background information on their year of birth, country, city, gender, and their preferred learning method (none, abroad, online, near home town). Test takers' first language, age of onset, or highest education completed, for example, were not asked for. We do not have the individual test takers' scores and background information.[1] The SET scores are only available on country level, for 112 countries and regions.

Description of the data

EF (2021) reports that 53% of the language learners was female, 96% was below 60 years of age, and their median age was 26 years. Data of at least 400 language learners were available for each city, region, and country included in the EF (2021) report. Many countries in Africa and Asia are former British colonies where English nowadays is an official language or medium of instruction in secondary and tertiary education. We have included that information in our data file on the base of information provided by World Atlas (2022). In addition, we have included first languages (L1s) of the test takers on the basis of their country. We, again, used information provided by World Atlas (2022) and selected a country's official indigenous language most widely spoken. Finally, we included three types of linguistic distances for each

[1] We have tried to get access to individual data, not just of EF, but of ETS and Cambridge, as well. Yet, these test institutes are found to be very reluctant to provide such data. We, of course, appreciate their concern on the sensibility of data at the test item level. On the other hand, it is evident that such test data, and even overall test scores, could provide a goldmine for SLA researchers. A notable exception is the State Exam Dutch as an L2, which has provided us the overall test scores on speaking, writing, listening and reading along with background info of learners of Dutch as an L2.

first language involved with respect to English. These types of linguistic distance are: 1) lexical distance, 2) morphological distance, and 3) phonological distance (see below). We were able to determine linguistic distance scores for 61 of the 63 first languages. For Haitian creole (Haiti) and Kinyarwanda (Rwanda) we were unable to determine their morphological distance score because an insufficient number of relevant morphological features are available (Dryer & Haspelmath, 2011).

Variables

Standard English Test score, aggregated scores per country (*N* = 110, *M* = 508, *SD* = 70).

Human Capital Index score, "the HCI measures the *expected* future human capital of a child born today, given *current* education and health outcomes for the young" (Kraay 2018, p.4). "The HCI consists of three components: (1) survival, measured as the probability of survival to age five; (2) school, which combines a measure of the number of years of school a child born today can expect to attain given prevailing enrollment rates with a measure of the quality of education based on international student achievement tests; and (3) health, which uses childhood stunting rates and adult survival rates as proxies for the overall health environment" (Kraay, 2018, p.2). The HCI is available on country level (*M* = 0.59, *SD* = 0.13).

Status of English (SE), whether (1) or not (0) a country uses English as one of the official languages, next to an indigenous language or as a medium of instruction in secondary schooling (M = 0.23, SD = 0.42). This information was obtained from the World Atlas (2022).

First language (L1), a country's largest indigenous language. There are 61 languages present. These include 30 Indo-European languages, five Afro-Asiatic, five Altaic, four Austronesian, three Atlantic-Congo, three Sino-Tibetan, three Uralic, two Austro-Asiatic, and one Dravidian, Japonic, Kartvelian, Korean, Nilotic, and Tai-Kadai language. We have collapsed the language families that contain just one or two languages (eight in total) into a category assigned as "Other". Two L1s occur by far the most frequent: Arabic (*N* = 17*)* and Spanish (*N* = 19).

Indo-European language (IE), whether (1) or not (0) a first language belongs to the Indo-European language family. English belongs to the Indo-European language family. We added this variable to ascertain that our linguistic distance measure is not biased towards the Indo-European language family or that this distinction may be even stronger than the linguistic distance measure. We, therefore, want to test if a simple distinction between Indo-European or not is sufficient to capture the efficiency factor (*M* = .54, *SD* = .50).

Lexical distance, a symmetric measure that represents the sum of branch lengths that connect two languages in a phylogenetic language tree of the Indo-European language family (see: Schepens et al., 2013b for the Dutch version of this measure). This measure is assumed to be particularly sensitive for distances between English and other Indo-European languages and the consequence is a maximum distance between English and non-Indo-European languages are 19.03, which also is the score of Urdu, the lexically most distant Indo-European language to English (*M* = 15.19, *SD* = 4.18).

Morphological distance, an asymmetric measure that compares the morphological features between languages according to differences in complexity (see: Schepens et al., 2013a for the Dutch version of this measure). We used an existing list with rank orderings for the feature values of 28 morphological features (Lupyan & Dale, 2010). We computed distances for the 61 languages that have at least five available values in WALS (Dryer & Haspelmath, 2011). This measure is assumed to be particularly sensitive for distances to non-Indo-European languages (M = 0.24, SD = 0.12). We redefined one of the morphological features, feature 77 which stands for "Semantic distinctions of evidentiality" with feature values: 1) no grammatical evidentials, 2) indirect only, and 3) direct and indirect. In previous research on morphological distance to Dutch, we made a distinction between 1 versus 2 and 3. This turned out to be a bad decision of English as the target language, as English is the only Germanic language with feature value 1. We therefore decided to make a distinction between 1 and 2 versus 3.

Phonological distance, an asymmetric measure counting the number of new phonological features in a target language based on complete sound and feature inventories (Schepens et al., 2020). The phonological sound and feature inventories from PHOIBLE (Moran & McCloy, 2019) were used. We computed distances to English for the 61 languages for which PHOIBLE lists a phoneme inventory (M = 13.85, SD = 4.07).

Statistical approach

To answer or first two questions, we conducted correlational analyses and principal component analysis. For answering our third question we applied linear mixed-effects analysis by using the 'lme4' package (Bates et al., 2015) in R (R Core Team, 2018). We present additional outcomes and analyses in the Supplementary Material.

Results

The linguistic distance factor (RQ 1)

In order to answer the first question, we computed correlations on the level of the 61 languages, between the three distance measures and the SET scores. When languages were present in more countries, we computed their mean score. We made a distinction between Indo-European and non-Indo-European languages. Figure 1 gives the correlations and scatterplots between the four variables included in our analysis. The diagonal contains the density plots of their distributions.

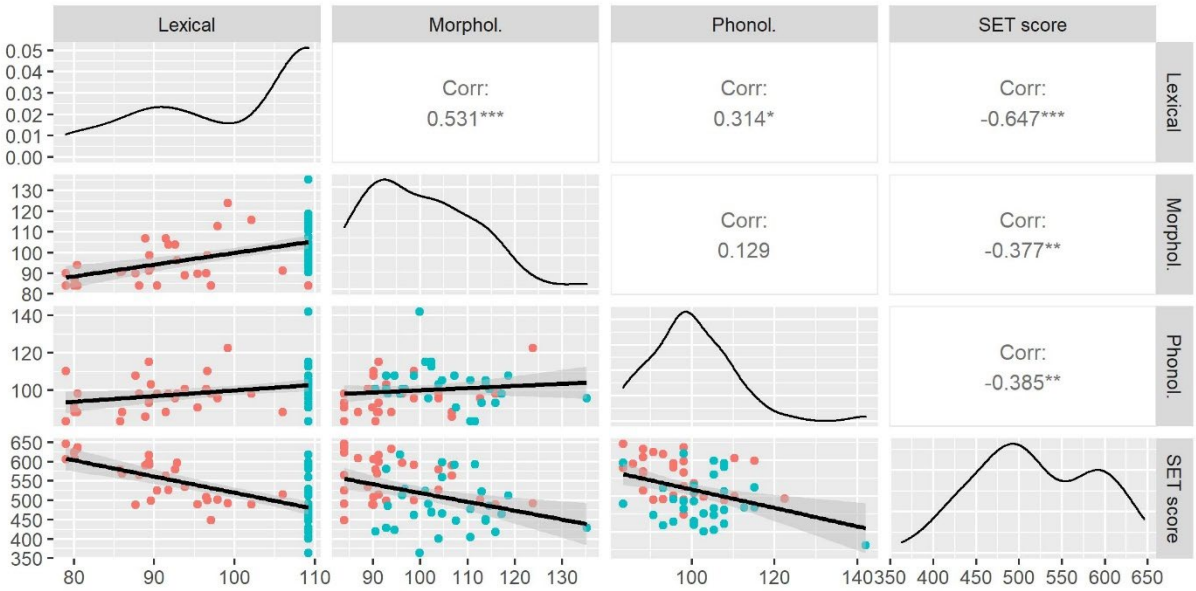

Figure 1: Correlations between lexical, morphological, phonological distance measures and the mean SET country scores of each language (30 IE-languages in pink, 31 non-IE languages in mint)).

As can be seen in Figure 1, all three distance measures are correlated significantly with the SET scores in the expected direction: the more distant the language is, linguistically, the lower its SET score is. Clearly, the correlation with the lexical distance measure is the strongest, while the morphological and phonological distance measures are both mediocre in magnitude (r = .377, p < .010 and r = .385, p < .010, respectively). It can also observed that lexical distance correlates substantially (r = .531, p < .001) with morphological distance. As for the lexical distance, the non-IE language all have a maximum dissimilarity, because they don't share any cognate. The morphological and phonological measures have mixed patterns.

Given three significant correlations of the three distance measures involved with the SET scores, we decided to bring them together in one comprehensive *linguistic distance measure*. In order to do so, we performed a principal component analyses (R Core Team, 2018). The outcomes are displayed in Table 1.

Table 1: PCA, one-factor solution, extraction criterion: lambda > 1 (k = 61 languages)

	Linguistic distance	h^2
Lexical distance	.870	.757
Morphological distance	.786	.618
Phonological distance	.552	.305
SS loadings (eigen value)		1.679
Prop. variance		.560

Clearly, the three linguistic distance measures can be subsumed in one linguistic distance measure (LD). The proportion variance covered by this factor is .56, which may be considered as substantial. We calculated factor scores based on the Maximum Likelihood method and added them to the data base, (M = xx, SD = yy).

The correlations between the factors and the EF SET country scores (RQ2)

Figure 2 gives the correlations and scatterplots between the four variables or determinants that we included in our data analysis. The diagonal contains the density plots of their distributions.

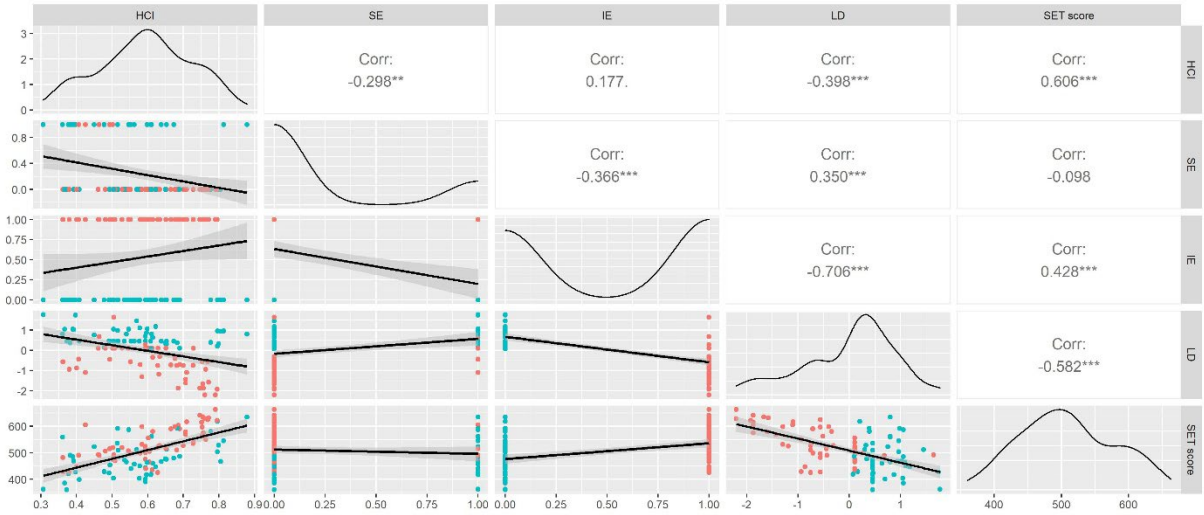

Figure 2: Correlations between the HCI, SE, IE, and LD determinants and the SET country scores (59 IE-countries in pink, 51 non-IE countries in mint))

The human capital index (HCI) has the highest correlation with the SET score (0.606), but the correlation of the linguistic distance measure (LD) is not much lower in strength (-0.582). Whether or not coming from a country with an Indo-European language (IE) has a significant correlation, but the English status of a country (SE) has not. There are also correlations between the factors. That means that we have to apply regression analyses to investigate in which way our factors complement each other or not.

Modeling all determinants (RQ3)

Table 2 describes four successive models in a stepwise forward selection process by adding new determinants, with the final Model 3 comprising four determinants. All these predictors were centred around their grand mean to reduce multicollinearity, if present. We started with a base model, Model 0, containing only a random effect for language. After adding HCI and SE as determinants in model 1. The contribution if SE is not significant. We proceeded by adding the Indo-European (IE) factor in model 2. This determinant is significant, SE becomes also significant, LD is added in model in model 3. The balance between the other three determinants shifts, HCI being a bit weaker, but SE is again significant, whereas IE is now no longer significant. Another relevant outcome is the total error variance that is the sum of the residual variance and the (remaining) L1variance. The total amount clearly reduces, the L1 variance sharply reducing when the linguistic determinants, IE and LD, are being included. It means that the initial random variation between languages has a systematic component that can be explained by linguistic factors.

Table 2: Multilevel model parameter estimations of the EF Standard English Test (standard errors in parentheses), [standard deviations in brackets] per first language (110 countries, 61 L1s). Independent variables are standardized

Model	Model 0	Model 1	Model 2	Model 3
Intercept	519*** (8.8)	515*** (7.1)	517*** (6.1)	515*** (5.7)
HCI		39.57*** (5.28)	38.09*** (4.92)	34.03*** (5.081)
SE		6.92 (5.08)	12.02** (4.87)	13.95** (4.80)
IE			29.01*** (6.25)	10.89 (9.10)
LD				– 22.54** (8.63)
Variance components				
Residual	2,328[48]	1,618[40]	1,613[.40]	1,641[41]
Variance L1	2,782[53]	1,707[11]	935[31]	683[26]
REML	1,216	1,160	1,136	1,123

*: $p < .05$; **: $p < .01$; ***: $p < .001$

To evaluate our models in more detail, we calculated Nakagawa's *conditional* and *marginal* R^2*s* (Nakagawa et al., 2017) using the 'performance' R package (Lüdecke et al., 2020) for the EF SET scores and each of the four models, see Table 3. In addition, we give four fit measures in Table 3. All measures point to Model 3 as the best model.

Table 3: Model improvement statistics for EF SET scores

Model	*df*	*AIC*	*BIC*	*deviance*	$-\Delta Chi^2$	*Δdf*	*P*	*Marginal* R^2	*Conditional* R^2
Model 0	3	1228	1236	1222	0			0	.544
Model 1	5	1186	1199	1176	46.4	2	< .001	.304	.661
Model 2	6	1169	1185	1157	19.3	1	< .001	.474	.667
Model 3	7	1164	1183	1150	6.8	1	= .009	.513	.656

We evaluated the quality of the final model by means of checking a variety of model assumptions. The outcomes can be found in the Supplementary Material. The main conclusion to be drawn from the model assumptions check is that Model 3 is our final model. We added in subsequent models interaction between the determinants, but none of these were significant or increasing the model fit.

In Figure 3, we present a graphical presentation of the strength of the effects and their confidence intervals, using the package 'sjPlot' (Lüdecke, 2022). IE includes the 0 value,

meaning that this predictor is not significant. The remaining three effects are the determinants according to the theoretical model of Chiswick and Miller (1995, 2014).

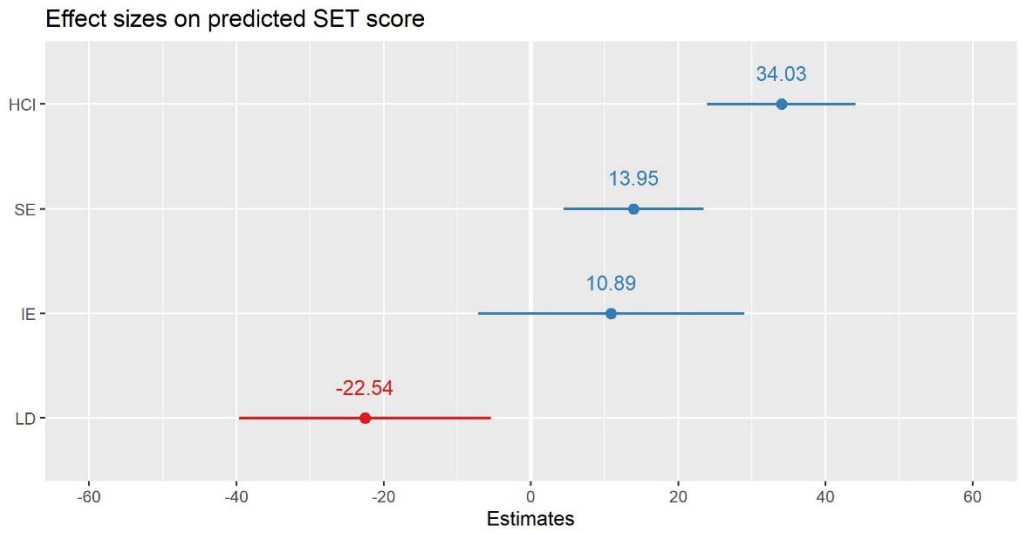

Figure 3: Effect sizes and confidence intervals (95%) of, downwards, the HCI, SE, IE and LD predictors on the predicted SET score

Conclusion and Discussion

Chiswick and Miller (1995, 2014) developed an economic model for proficiency in the destination language of immigrants. The language in question was English and they view immigrant language skills as a part of human capital. We applied the three core concepts of their economics of language model again, but now to English as a global language. The SET country data is largely based on people doing the test in their home country or outside an English only speaking country. We added three new elements. We used real country proficiency scores based on real test data, and not self-reported language skills. Secondly, we used the Human Capital Index (HCI) to measure the economic status and potential of a country or region. Lastly, we introduced an objective linguistic distance measure. This relates to our first research question how to operationalize linguistic distance.

Chiswick and Miller (2005) had to resort to reported difficulties of native speakers of English learning other languages. Chiswick and Miller (2014) positively and extensively refer to later studies where linguistic distance measures were used, based on language trees. In our earlier work (see Schepens et al. forthcoming), we developed linguistic measures in a new way, not only based on words, but on constructs and sounds as well. Our original algorithms were tuned to Dutch as the destination language, but we adapted them to compute distances of other languages in relation to English as the target language (see Van der Slik et al. 2019). We combined them in one overall measure, which correlated well with the SET country scores. In addition, its application is not restricted to a language tree. Language trees have the restriction to be applicable only within a language family, the most famous one being the Indo-European language, with English as one of its members. The consequence is that all languages of other families have the same maximal distance. We added a morphological and a phonological measure to overcome that problem and constructed one overall distance measure, that has a

high correlation with the SET country scores. We included the distinction between Indo-European and non-Indo-European in our analyses to ascertain that our linguistic distance measure was superior, as compared to a crude distinction between Indo-European and non-Indo-European languages.

The SET country score was our criterion variable. Unfortunately, we had no access to individual learner data to assess, for instance, the effects of gender, age, and education. EF (2021) reports that for the first time in their SET testing history, male learners' SET scores surpassed those of female learners slightly. In a previous study on Dutch as a destination language (Van der Slik, et al, 2015) we concluded that female learners of Dutch were outperforming male learners, but we also observed systematic gender differences across countries. How valid or representative are those country scores? EF illustrates their value by reporting each year their correlation with economic indices, including HCI. TOEFL, another global player in testing English language skills, also reports country scores, adding year after year, the following statement: "ETS, creator of the TOEFL® test, does not endorse the practice of ranking countries on the basis of TOEFL scores, as this is a misuse of data. The TOEFL test provides accurate scores at the individual level; it is not appropriate for comparing countries" (TOEFl iBT 2020, p. 19). We decided to use the SET country scores. Despite TOEFL's statement, their 2019 country and region scores have a correlation of .81 with the SET country scores (EF, 2021, p.32). This high correlation in fact corroborates the validity of the data we used.

Because of the lack of individual data, we had to generalize over the mother tongues of the learners. For monolingual countries, this is fairly unproblematic. There is of course some noise due to expats' or immigrants' L1s. Most countries are multilingual however. Nigeria and Cameroon, for example, can be qualified form a linguistic point of view as paradises with more than 500 and 200 first languages, respectively (World Atlas, 2022). We decided to choose Hausa and Fulfulde as first languages for Nigeria and Cameroon. Fulfulde was chosen because it is a lingua franca in the north of Cameroon. Alternatively, Ewondo might have been selected as it is a lingua franca in East, South and center of Cameroon. For Nigeria, we have selected Hausa because it is the most frequent spoken indigenous language. Alternatively, Yoruba or Igbo might have been chosen, as well. Also, South Africa is a multilingual country with eleven official languages of which Zulu, Xhosa, Afrikaans, and English in descending order are most widely spoken. Our choice of first languages was also guided by the availability of relevant morphological features in the World Atlas of Language Structures. For that reason, we have decided to pick Tamil instead of Sinhalese as L1 in Sri Lanka, for Kazakhstan and Belarus we have chosen Russian instead of Kazakh and Belarussian, for the Philippines: Tagalog instead of Cebuano, in Luxembourg, German instead of Luxembourgish, and Dinka in South Sudan instead of Bari. For two countries we were unable to select a first language with adequate information on their morphological features (Haitian Creole in Haiti and Kinyarwanda in Rwanda). Our choices did not have severe consequences for our results. We found no (influential) outliers in our data analysis in the languages and countries.

Our three core determinants of the SET country scores were linguistic distance (LD), the status of English (SE), and the Human Capital Index (HCI). These correlations were significant (answering research question 2), and we made them part of mixed regression analyses with language as a random factor (to answer research question 3). Our final model

which included these three determinants had a high, persuasive level of explained variance (51%). It means that in an economic model of English as a global language, linguistic distance is a crucial cost factor, related to the concept of efficiency. The more distant the home language the more effort is required to become proficient in English. The status of English relates to exposure. The Human Capital Index is a mixture of the concepts of efficiency and exposure. Additionally, we assume that learning English has a number of economic incentives for the learners. It gives them, amongst other opportunities, access to international networks, jobs and markets.

How can we compare the economic models of English as an immigrant language and English as a global language? They seem to be based on the same concepts, but the intensity and strength of their determinants may work out differently. Hartshorne, Tenenbaum, and Pinker. (2018) collected grammatical judgment data on English of almost one million participants. These data reflect proficiency in English. A clear outcome was that immersion learners (immigrants) had higher scores than non-immersion learners (English as a global language), a distinction that they attribute to experience (the "experience discount factor") a concept that can be subsumed under exposure. Most economic studies used self-reported or -assessed language proficiency measures. Comparing the two economic models would profit from using the same proficiency measures, preferably objective measures, like the SET test data.

We have demonstrated that the application of an objective linguistic distance measure makes sense. We need more data to develop this measure further, by collecting information on its balance between the three components: words, constructs, and sounds. We discuss this in Schepens et al. (forthcoming). The linguistic possibilities are increasing rapidly, given the extensive linguistic databases that are becoming available.

Literature

Azoth Analytics (2019). *Global English proficiency test market: Insights, trends and forecast (2019-2024).*

Chiswick, B.R., & Miller, P.W. (1995). The endogeneity between languages and earnings: International analyses. *Journal of Labor Economics, (2)13,* 246 – 288.

Chiswick, B.R. , & Miller, P.W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development, 26 ,* 1 – 11.

Chiswick, B.R., & Miller, P.W. (2014). International migration and the economics of language. *IZA Discussion Paper No. 7880*. http://dx.doi.org/10.2139/ssrn.2381132

Dryer, M.S., & Haspelmath, M. (Eds.). (2011). *The world atlas of language structures online. Max Planck Digital Library.* http://wals.info/

Dutta, S., Lanvin, B, & Wunsch-Vincent, S. (eds.) (2020). *Global innovation index 2020. Who will finance innovation?* World Intellectual Property Organization. https://www.globalinnovationindex.org/analysis-indicator Retrieved at 31-01-2022.

Education First (2014). *EF SET Academic and technical development report.* https://www.efset.org/about/ Retrieved at 20-01-2022.

Education First (2015a). *EF SET PLUS – IELS correlation study report.* https://www.efset.org/about/ Retrieved at 20-01-2022.

Education First (2015b) *EF SET PLUS – TOEFL iBT correlation study report.* https://www.efset.org/about/ Retrieved at 20-01-2022.

Education First (2021) *EF EPI EF English Proficiency Index. A ranking of 112 countries and regions by English skills.* https://www.ef.nl/epi/ Retrieved at 20-01-2022.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature, 426(6965),* 435–439. https://doi.org/10.1038/nature02029

Hartshorne, J.K., Tenenbaum, J.B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177,* 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Kraay, A.(2018). The World Bank human capital index: A Guide. *The World Bank Research Observer, (1)34*, 1-33.

Lanvin, B., & Monteiro, F. (eds.) (2020). *The global talent competitiveness index 2020. Global talent in the age of artificial intelligence* INSEAD. https://www.insead.edu/faculty-research/publications/reports/the-global-talent-competitiveness-index-2020-global-talent-in-the-age-of-artificial-intelligence-40131

Lüdecke D (2022). *sjPlot: Data visualization for statistics in social science*. R package version 2.8.10.1, https://CRAN.R-project.org/package=sjPlot.

Lüdecke, D., Makowski, D., & Waggoner, P. (2020). *Performance: Assessment of regression models performance (0.4.4) [Computer software].* https://CRAN.R-project.org/package=performance

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE, 5(1),* e8559. https://doi.org/10.1371/journal.pone.0008559

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0.* Max Planck Institute for the Science of Human History. https://phoible.org/

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14(134),* 20170213. https://doi.org/10.1098/rsif.2017.0213

Nieuwenhuis, R., Te Grotenhuis, H. F., & Pelzer, B. J. (2012). Influence. ME: Tools for detecting influential data in mixed effects models. *The R-Journal, 4(2),* 38–47. https://doi.org/10.32614/RJ-2012-011

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Schepens, J. (2015) *Bridging linguistic gaps: The effects of linguistic distance on the adult learnability of Dutch as an additional language.* Published PhD Thesis. Utrecht: LOT.

Available at: www.lotpublications.nl/Documents/383_fulltext.pdf (accessed January 2017).

Schepens, J., Van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition, 194,* 104056. https://doi.org/10.1016/j.cognition.2019.104056

Schepens, J., Van der Slik, F., & Van Hout, R. (forthcoming). Linguistic dissimilarity increases cognitive aging effects in adult language learning. *Studies in Second Language Acquisition*

Schepens, J.J., Slik, F.W.P. van der & Hout, R.W.N.M. van (2016). L1 and L2 distance effects in learning L3 Dutch. *Language Learning,* 66(1): 224-256. https://doi.org/10.1111/lang.12150

Takeno, J. & Moritoshi, P. (2018). Re-examining the English proficiency level of Japanese EFL learners. *Chugokugakuen Journal, 17,* 35-39.

Tibus, E.. & Bendulo, H.O. (2018). Educational, economic, and employment influences to English language proficiency. *Journal of Educational and Human Resource Development, 6,* 274-279.

United Nations Conference on Trade and Development. (2020). *UNCTAD Productive capacities index. Methodological approach and results.* *https://unctad.org/webflyer/unctad-productive-capacities-index-methodological-approach-and-results* Retrieved at 20-01-2022.

Van der Slik, F. (2010). Acquisition of Dutch as a second language. *Studies in Second Language Acquisition, 32(03),* 401–432. https://doi.org/10.1017/S0272263110000021

Van der Slik, F., Van Hout, R., & Schepens, J. (2015). The gender gap in second language acquisition: Gender differences in the acquisition of Dutch among immigrants from 88 countries with 49 mother tongues. *PLoS ONE, 10(11),* e0142056. https://doi.org/10.1371/journal.pone.0142056

Slik, F. van der, Hout, R. van & Schepens, J. (2019). The role of morphological complexity in predicting the learnability of an additional language. The case of La (additional language) Dutch. *Second Language Research,* 35(1): 47-70. https://doi.org/10.1177/0267658317691322

World Atlas (2022). https://www.worldatlas.com/articles/what-languages-are-spoken-in-singapore.html Retrieved at 20-01-2022

World Bank. (2020). *Data Bank | Human Capital Index.* https://databank.worldbank.org/source/human-capital-index Retrieved at 20-01-2022

World Economic Forum (2022). https://www.weforum.org/agenda/2019/11/countries-that-speak-english-as-a-second-language/ Retrieved at 21-01-2022

Supplementary Material

1. Validity of the EF SET scores
2. Quality check of model assumptions
3. Random intercepts

Validity of the EF SET scores

In order to get an indication on how valid EF SET scores are, we correlated these scores with the five scores on TOEFL 2020 test (ETS, 2022). TOEFL presents both scores on the L1 as the country level. This allows us to calculate correlations on both levels, see Table S1.

Table S1: Correlations of the EF SET country and L1 scores with the 2020 TOEFL speaking, writing, reading, and listening scores on both the country and the L1 level[#]

	Speaking	Writing	Reading	Listening	All
SET scores (57 L1s)	.65	.66	.63	.71	.71
SET scores (109 Countries)	.70	.78	.74	.79	.80

[#] all correlations at $p < .001$

As EF (2021) only presents SET scores at the country level, we had to deduce a country's dominant language from information provided by World Atlas (2022). Reviewing the correlations provided in Table S1, we come to the conclusion that this deduction has not been without success. It should be noted however that ETS explicitly and repeatedly does not endorse the use of their data by comparing L1 and country scores as we have done here. We nevertheless think it is encouraging to find such high correlations between the EF SET and TOEFL, from our point of view.

Quality check of model assumptions

We have checked the quality of the final model by means of the following analyses: normality of SET scores and residuals, and influential cases.

Normality check

Figure S1 depicts that SET scores are distributed normally, while Figure S2 shows only a marginal departure from normality of residuals for Model 3.

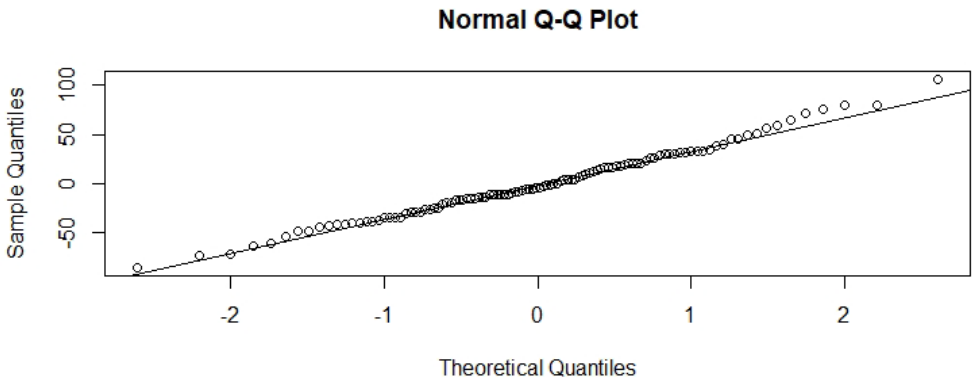

Figure S1: Distribution of the SET scores of Model 3

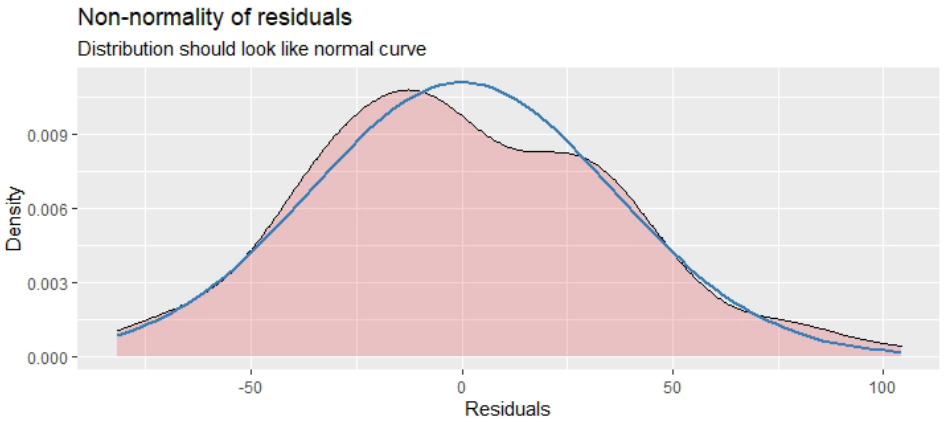

Figure S2: Distribution of the residuals of Model 3

Check on influential cases

In order to check if the results are affected by the occurrence of influential cases, we applied a dfBeta analysis (Nieuwenhuis, Te Grotenhuis, & Pelzer, 2012). Figure S3 presents the results. The only potential influential case found is Arabic. English language learners with Arabic as L1 seem to score lower as expected given their human capital index score and whether or not their country has had historical ties with Great Brittan. We have checked if results would change if we compare the outcomes of the final model with and without Arab-speaking countries by means of the 'EMAtools' package (Kleiman, 2017). This proved not to be the case as the parameter estimates of both models do not differ significantly.

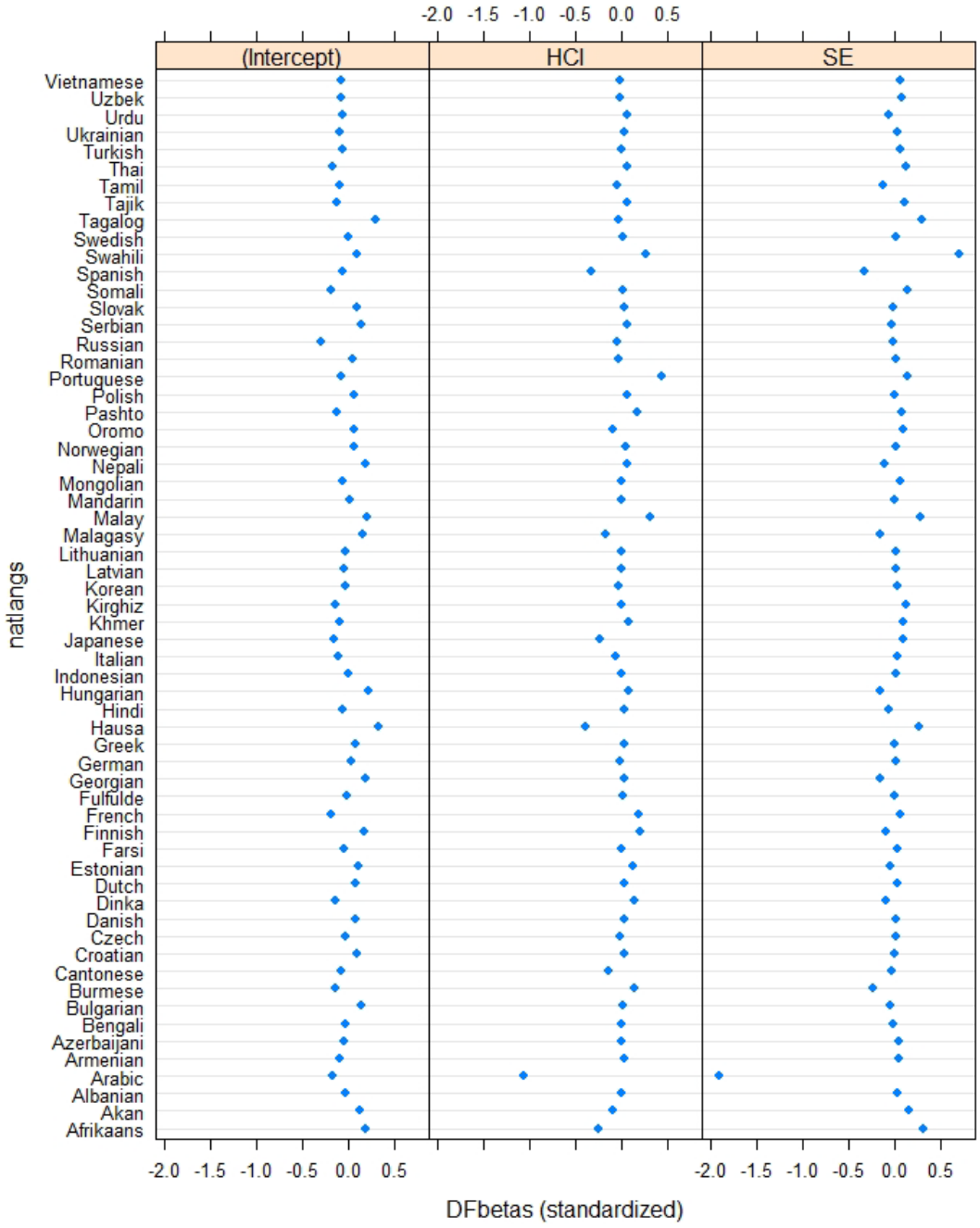

Figure S3: Influential cases (DFbetas) against Human Capital Index scores (HCI) and status of English (SE)

Random intercepts

In Figure S4, we present the random intercepts of Model 3. It seems clear that the 95% CI's are quite large which is of course to be expected as most L1s occur only once. There are two cases of interest. These are the L1s Russian and Arabic as both score significantly below mean. It is difficult to give a reason for these low scores. One might speculate that some unmeasured factors are responsible for these divergent scores, but we refrained of doing so as we first want to have more evidence, for instance by including more testing years.

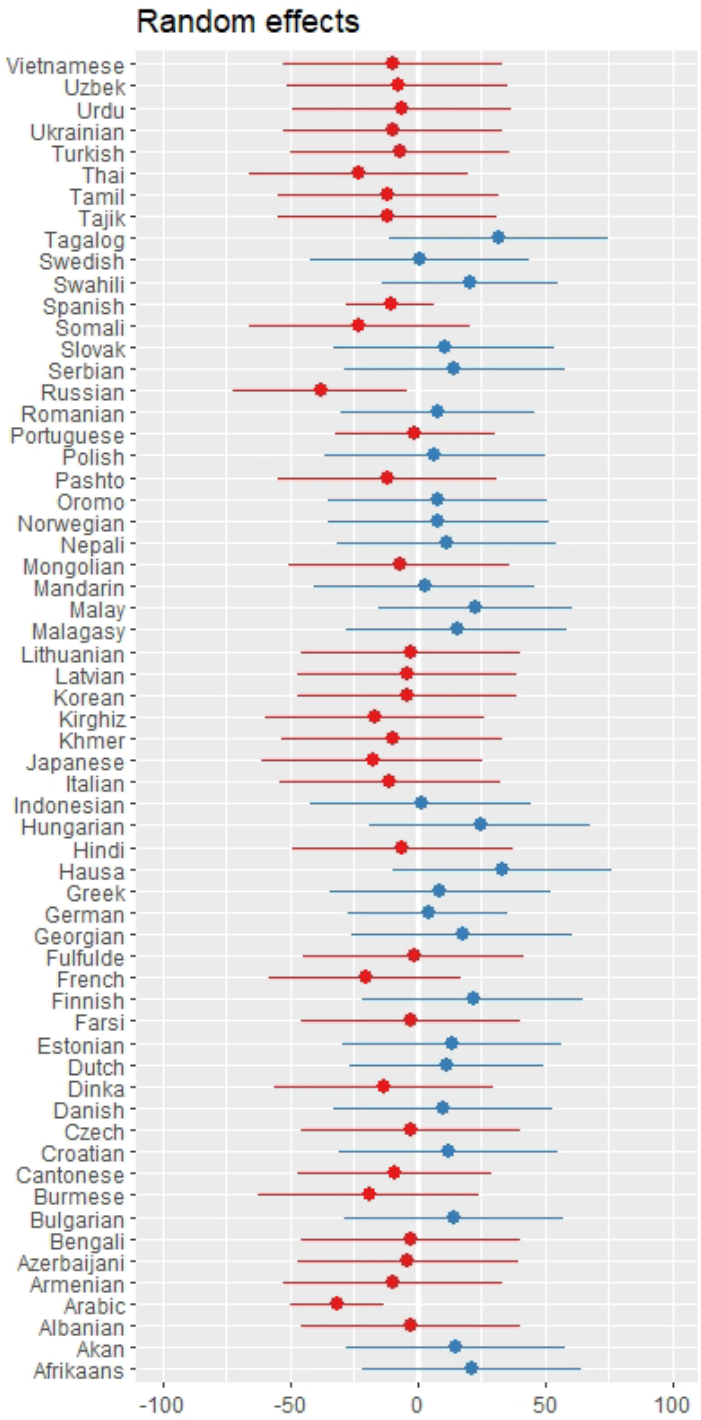

Figure S4: Random effects of Model 3